ON DETECTING INFLUENTIAL DATA AND SELECTING REGRESSION VARIABLES*

by

Shanti S. Gupta
Purdue University

and

Deng-Yuan Huang
National Taiwan Normal University

Technical Report #89-28C

Department of Statistics
Purdue University

December 1989

* This research was supported in part by the Office of Naval Research Contract N00014-88-K-0170 and NSF Grants DMS-8606964, DMS-8702620 at Purdue University.

ON DETECTING INFLUENTIAL DATA AND SELECTING REGRESSION VARIABLES

by

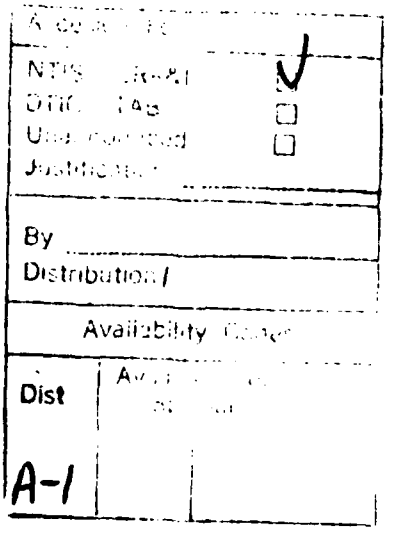
Shanti S. Gupta
Purdue University

Deng-Yuan Huang
National Taiwan Normal University

Abstract

We consider a linear regression model. We attempt to measure the influence on the residual for the full model and the reduced model. A criterion to select important independent regression variables is also derived using the influence diagnostics. This criterion turns out to be the same as the one proposed by Gupta and Huang (1988).

Key Words: Linear Model, Influential data, Selection criteria, Inferior model.



On Detecting Influential Data and Selecting Regression Variables*

by

Shanti S. Gupta
Purdue University

Deng-Yuan Huang
National Taiwan Normal University

1. Introduction

We consider the following linear model

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon}, \tag{1}$$

where $\underline{\varepsilon} \sim N(\underline{0}, \sigma_0^2 I_n)$, I_n denotes the identity matrix of order n, $\underline{Y}$ is an $n \times 1$ vector of responses, X is a $n \times p$, $(n > p)$, matrix of known constants of rank p, $\underline{\beta}$ is a $p \times 1$ parameter vector. Several authors have studied the influence on the fitted regression line when the data are deleted. Let $\underline{\hat{\beta}}$ be the usual least squares estimator of $\underline{\beta}$ based on the full data and let $\underline{\hat{\beta}}_A$ be an alternative least squares estimator based on a subset of the data. The empirical influence function for $\underline{\hat{\beta}}$, IF_A is defined to be

$$IF_A = \underline{\hat{\beta}}_A - \underline{\hat{\beta}}. \tag{2}$$

For a given positive definite matrix M and a nonzero scale factor c, Cook and Weisberg (1980) defined the distance $D_A(M, c)$ between $\underline{\hat{\beta}}$ and $\underline{\hat{\beta}}_A$ as follows:

$$D_A(M, c) = \frac{(IF_A)'M(IF_A)}{c}. \tag{3}$$

Cook and Weisberg (1980) suggest that the matrix M can be chosen to reflect specific interests.

Cook and Weisberg (1980) pointed out that in some applications, measurement of the influence of cases on the fitted values, $\underline{\hat{Y}} = X\underline{\hat{\beta}}$, may be more appropriate than measuring influence on $\underline{\hat{\beta}}$. They mentioned an example to describe the fact that if prediction is the primary goal it may be convenient to work with a reparameterized model where the regression coefficients are not of interest. Cook and Weisberg (1980) tried to treat their measurement of the influence on the fitted values $X\underline{\hat{\beta}}$. They used the empirical influence function for $\underline{\hat{Y}}$ as defined by $X(IF_A)$. In this paper, we attempt to measure the influence

* This research was supported in part by the Office of Naval Research Contract N00014-88-K-0170 and NSF Grants DMS-8606964, DMS-8702620 at Purdue University.

on residuals or on $X\underline{\hat{\beta}}$. The large influence on the residual should have much influence on $\underline{\hat{\beta}}$ though the converse may not hold. Furthermore, Welsch (1982) pointed out that in an earlier paper Cook (1977) chose to measure influence by

$$D = \frac{(\underline{\hat{\beta}} - \underline{\hat{\beta}}_{(i)})' X' X (\underline{\hat{\beta}} - \underline{\hat{\beta}}_{(i)})}{s^2 p} \tag{4}$$

where s^2 is the residual mean square for full data. Welsch (1982) gave an example to explain that when all of the observations but one lie on a line, (4) can give potentially confusing information since it may indicate that some observations on the line are more influential than the one observation not on the line. This is counterintuitive since the deletion of this one observation leads to a perfect fit. Welsch (1982) gave the example as follows:

Let $x_1 = x_4 = 10$, $x_2 = 11$, $x_3 = -11$; $y_1 = 10$, $y_2 = 11$, $y_3 = -11$, $y_4 = 10.5$ and fit a simple linear regression model with intercept. The fitted simple regression line is

$$\hat{Y} = 0.0885 + 1.0073X,$$

$$R^2 = 0.9995, \quad \text{Root MSE} = 0.2909, \quad \text{C.V.} = \frac{100(\text{Root MSE})}{\overline{Y}} = 5.68\%,$$

and the values of Cook's D are: (1) 0.109, (2) 0.144, (3) 178.495, (4) 0.477. We find that the observation 3 is much more influential than observation 4. Therefore, finding a more reasonable measurement is very important. We shall consider the case of one data deletion at one time. Since, for the deletion of any subset case, computations can be similarly carried out, we refer to Cook and Weisberg (1980), Gray and Ling (1984).

Next we propose a selection criteria to combine the influence measure and variable selection. We derive a suitable choice of M and c in (3) to measure the influence and bias for the reduced model. Then, the inferior reduced model can be determined. An example (Daniel and Wood (1980)) is studied to explain the idea for the proposed criteria.

2. Influential Observations in Linear Regression Model

Let $\underset{n\times p}{X} = \binom{X_{(i)}}{X_i'}_{1\times p}^{(n-1)\times p}$, $\underset{n\times 1}{\underline{Y}} = \binom{\underline{Y}_{(i)}}{Y_i}_{1\times 1}^{(n-1)\times 1}$, $\underline{\hat{Y}} = X\underline{\hat{\beta}}$, $\underline{e} = \binom{\underline{e}_{(i)}}{e_i}_{1\times 1}^{(n-1)\times 1}$, $e_i =$

$Y_i - \underline{X}_i'\underline{\beta}$, $i = 1, 2, \ldots, n$, $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{Y}$. Then

$$\begin{aligned}
\underline{e}_{(i)}'\underline{e}_{(i)} &= (\underline{Y}_{(i)} - X_{(i)}\underline{\beta})'(\underline{Y}_{(i)} - X_{(i)}\underline{\beta}) \\
&= (\underline{Y}_{(i)} - X_{(i)}\hat{\underline{\beta}}_{(i)} + X_{(i)}\hat{\underline{\beta}}_{(i)} - X_{(i)}\underline{\beta})'(\underline{Y}_{(i)} - X_{(i)}\hat{\underline{\beta}}_{(i)} \\
&\quad + X_{(i)}\hat{\underline{\beta}}_{(i)} - X_{(i)}\underline{\beta}) \\
&= (\underline{Y}_{(i)} - X_{(i)}\hat{\underline{\beta}}_{(i)})'(\underline{Y}_{(i)} - X_{(i)}\hat{\underline{\beta}}_{(i)}) \\
&\quad + (X_{(i)}\hat{\underline{\beta}}_{(i)} - X_{(i)}\underline{\beta})'(X_{(i)}\hat{\underline{\beta}}_{(i)} - X_{(i)}\underline{\beta}),
\end{aligned}$$

Since $(\underline{Y}_{(i)} - X_{(i)}\hat{\underline{\beta}}_{(i)})'X_{(i)} = \underline{Y}_{(i)}'X_{(i)} - \hat{\underline{\beta}}_{(i)}'X_{(i)}'X_{(i)} = 0$. Thus

$$\underline{e}_{(i)}'\underline{e}_{(i)} = SSE_{(i)} + R_{(i)},$$

where

$$\begin{aligned}
SSE_{(i)} &= (\underline{Y}_{(i)} - X_{(i)}\hat{\underline{\beta}}_{(i)})'(\underline{Y}_{(i)} - X_{(i)}\hat{\underline{\beta}}_{(i)}) \\
R_{(i)} &= (\hat{\underline{\beta}}_{(i)} - \underline{\beta})'X_{(i)}'X_{(i)}(\hat{\underline{\beta}}_{(i)} - \underline{\beta})
\end{aligned}$$

and

$$SSE_{(i)} = \inf_{\underline{\beta}} \underline{e}_{(i)}'\underline{e}_{(i)}.$$

We have,

$$\frac{\underline{e}_{(i)}'\underline{e}_{(i)}}{SSE_{(i)}} = 1 + \frac{R_{(i)}}{SSE_{(i)}}. \tag{5}$$

Define

$$D_{(i)} = \frac{R_{(i)}}{ps_{(i)}^2}, \text{ where } s_{(i)}^2 = \frac{1}{n-p-1}SSE_{(i)}. \tag{6}$$

If $D_{(i)}$ is large, as in (5) we have that deleted i-th data will heavily influence the fitted line. We rewrite $R_{(i)}$ as follows:

$$\begin{aligned}
R_{(i)} &= (\hat{\underline{\beta}}_{(i)} - \underline{\beta})'X_{(i)}'X_{(i)}(\hat{\underline{\beta}}_{(i)} - \underline{\beta}) \\
&= [X_{(i)}(\hat{\underline{\beta}}_{(i)} - \underline{\beta})]'[X_{(i)}(\hat{\underline{\beta}}_{(i)} - \underline{\beta})] \\
&= (X_{(i)}\hat{\underline{\beta}}_{(i)} - X_{(i)}\underline{\beta})'(X_{(i)}\hat{\underline{\beta}}_{(i)} - X_{(i)}\underline{\beta}).
\end{aligned} \tag{7}$$

Thus, $R_{(i)}$ is the Euclidean distance between $X_{(i)}\underline{\beta}$ and its ordinary least square estimate (OLS) $X_{(i)}\hat{\underline{\beta}}_{(i)}$. Now, we use the full data to estimate $\underline{\beta}$ as the true value, we obtain the OLS estimate $\hat{\underline{\beta}}$.

We define a statistic to measure the influence in (5) for the fitted line as follows:

$$\hat{D}_{(i)} = \frac{\hat{R}_{(i)}}{ps^2_{(i)}} = \frac{(\underline{\hat{\beta}}_{(i)} - \underline{\hat{\beta}})' X'_{(i)} X_{(i)} (\underline{\hat{\beta}}_{(i)} - \underline{\hat{\beta}})}{ps^2_{(i)}}. \tag{8}$$

Let the hat matrix

$$H = \underset{n\times p}{X} (X' \underset{p\times p}{X})^{-1} \underset{p\times n}{X}{}' = (h_{ij})_{n\times n},$$

where

$$h_{ij} = \underline{X}'_i (X'X)^{-1} \underline{X}_j, \quad i,j = 1,\ldots,n.$$

We have

$$\begin{aligned}
(X'_{(i)} X_{(i)})^{-1} &= (X'X - X_i X'_i)^{-1} \\
&= (X'X)^{-1} + \frac{(X'X)^{-1} \underline{X}_i \underline{X}'_i (X'X)^{-1}}{1 - h_{ii}}, \\
\underline{X}'_i (X'_{(i)} X_{(i)})^{-1} \underline{X}_i &= \frac{h_{ii}}{1 - h_{ii}}, \qquad (9) \\
\text{Cov } (\underline{\hat{Y}}) &= X(X'X)^{-1} X' \sigma^2 = H\sigma^2, \\
\text{Var } (\hat{Y}_i) &= \underline{X}'_i (X'X)^{-1} \underline{X}_i \sigma^2 = h_{ii}\sigma^2, \quad i = 1,2,\ldots,n,
\end{aligned}$$

and

$$\text{Cov } (\underline{\hat{\beta}}) = (X'X)^{-1} \sigma^2.$$

Thus

$$\begin{aligned}
\underline{\hat{\beta}}_{(i)} &= (X'_{(i)} X_{(i)})^{-1} X'_{(i)} \underline{Y}_{(i)} \\
&= (X'X - X_i X'_i)^{-1} (X'\underline{Y} - \underline{X}_i Y_i) \\
&= \underline{\hat{\beta}} - \frac{(X'X)^{-1} \underline{X}_i \hat{e}_i}{1 - h_{ii}}, \text{ where } \hat{e}_i = Y_i - X'_i \underline{\hat{\beta}},
\end{aligned}$$

hence

$$\underline{\hat{\beta}}_{(i)} - \underline{\hat{\beta}} = -\frac{(X'X)^{-1} \underline{X}_i \hat{e}_i}{1 - h_{ii}}. \tag{10}$$

The i-th predict residual is

$$\hat{e}_{(i)} = Y_i - \underline{X}'_i \underline{\hat{\beta}}_{(i)}, \quad i = 1,2,\ldots,n. \tag{11}$$

Then

$$\begin{aligned}
\hat{e}_{(i)} = Y_i - \underline{X}'_i \underline{\hat{\beta}}_{(i)} &= Y_i - \underline{X}'_i \left(\underline{\hat{\beta}} - \frac{(X'X)^{-1} \underline{X}_i \hat{e}_i}{1 - h_{ii}} \right) \\
&= Y_i - \underline{X}'_i \underline{\hat{\beta}} + \frac{\underline{X}'_i (X'X)^{-1} \underline{X}_i \hat{e}_i}{1 - h_{ii}} \\
&= \hat{e}_i + \frac{h_{ii} \hat{e}_i}{1 - h_{ii}} = \frac{\hat{e}_i}{1 - h_{ii}},
\end{aligned} \tag{12}$$

and

$$\text{Var}\ (\hat{e}_{(i)}) = \frac{\text{Var}\ (\hat{e}_i)}{(1-h_{ii})^2} = \frac{\sigma^2}{(1-h_{ii})}.$$

Thus, we can obtain the following result,

$$\begin{aligned}\hat{R}_{(i)} &= (\underline{\hat{\beta}}_{(i)} - \underline{\hat{\beta}})' X'_{(i)} X_{(i)} (\underline{\hat{\beta}}_{(i)} - \underline{\hat{\beta}}) \\ &= [-\frac{(X'X)^{-1}\underline{X}_i \hat{e}_i}{1-h_{ii}}]' X'_{(i)} X_{(i)} [-\frac{(X'X)^{-1}\underline{X}_i \hat{e}_i}{1-h_{ii}}] \\ &= (1-h_{ii}) h_{ii} \hat{e}^2_{(i)}. \end{aligned} \tag{13}$$

Now

$$\hat{D}_{(i)} = \left\{ \left[\frac{\hat{e}_{(i)}}{s_{(i)} \cdot \frac{1}{(1-h_{ii})^{\frac{1}{2}}}} \right]^2 h_{ii} \right\} \frac{1}{p}. \tag{14}$$

The first factor in (14) is the studentized residual, that is, the residual divided by its standard error based on a fit to the data with the i-th case excluded, while the second term is the leverage of the variance of the i-th predicted value.

Since $\underline{\hat{\beta}}_{(i)}$ and $s^2_{(i)}$ are independent (cf. Graybill (1976)), and Y_i and $s^2_{(i)}$ are also independent. We have $\hat{e}_{(i)}$ and $s^2_{(i)}$ are independent. Thus

$$t_i = \frac{\hat{e}_{(i)}}{s_{(i)} \cdot \frac{1}{(1-h_{ii})^{\frac{1}{2}}}} = \frac{\hat{e}_{(i)}}{\frac{s_{(i)}}{\sigma} \cdot \frac{\sigma}{(1-h_{ii})^{\frac{1}{2}}}}$$

is t-distribution with $n-p-1$ degrees of freedom. Hence t_i^2 is F-distributed random variable with degrees of freedom 1 and $n-p-1$.

For given α, let C_i be satisfied the following equation:

$$P\{\hat{D}_{(i)} \geq C_i\} = \alpha.$$

Since $\hat{D}_{(i)} = t_i^2 h_{ii} \cdot \frac{1}{p}$, we have

$$P\{\hat{D}_{(i)} \geq C_i\} = P\{t_i^2 h_{ii} \cdot \frac{1}{p} \geq C_i\} = P\{t_i^2 \geq \frac{pC_i}{h_{ii}}\} = \alpha,$$

thus

$$\frac{pC_i}{h_{ii}} = F(1,\ n-p-1;\ 1-\alpha).$$

Hence

$$C_i = \frac{1}{p} h_{ii} F(1,\ n-p-1;\ 1-\alpha). \tag{15}$$

From equation (5):

$$\frac{\underline{e}'_{(i)}\underline{e}_{(i)}}{SSE_{(i)}} = 1 + \frac{\hat{R}_{(i)}}{SSE_{(i)}} \tag{16}$$

The estimated value of $\frac{\underline{e}'_{(i)}\underline{e}_{(i)}}{SSE_{(i)}}$ is:

$$\begin{aligned} 1 + \frac{p}{n-p-1}\hat{D}_{(i)} &\geq 1 + \frac{p}{n-p-1} \cdot \frac{1}{p} h_{ii} F(1,\ n-p-1;\ 1-\alpha) \\ &= 1 + \frac{1}{n-p-1} h_{ii} F(1,\ n-p-1;\ 1-\alpha). \end{aligned} \tag{17}$$

Thus, the i-th data is deleted, it will be the influential data, if $\hat{D}_{(i)} \geq C_i$. In the example (Table III) below, we find that the observation number $i = 29$ is an influential data. The amount of the influence for the residual will be at least $\frac{1}{n-p-1} h_{ii} F(1,\ n-p-1;\ 1-\alpha)$.

From (17), we define

$$IF_{(i)} = \frac{\hat{D}_{(i)}}{C_i}$$

as the measure of the strength of the influence on the residual when the i-th data is deleted.

We shall compare $\hat{D}_{(i)}$ with Cook's D in TABLE (I) and TABLE (II) for the Welsch (1982) example. We can find that $\hat{D}_{(4)}$ reflect the influence very seriously on the fitted line. If we deleted 4-th data, the line is fitted perfect. Hence, the $\hat{D}_{(i)}$ will be sensitively reflect the bias.

TABLE I

Obs.	Residual	RSTUDENT	HAT DIAG H
1	−0.1615	−0.5432	0.3231
2	−0.1689	−0.5948	0.3553
3	−0.0080	−0.5837	0.9985
4	0.3385	∞	0.3231

TABLE II

Obs.	$\hat{D}_{(i)}$	C_i	$IF_{(i)}$
1	0.0476	26.0810	0.0018
2	0.0629	28.6802	0.0022
3	0.1701	80.6001	0.0021
4	∞	26.0810	∞

3. Selecting Important Independent Variables

We shall consider the influential data in the reduced model. For the selection of important independent variables in (1), it is necessary to consider the measurement of the influence.

We denote model (1) as $\underline{Y}' = [Y_1, \ldots, Y_n]$, $X = [\underline{1}, \underline{X}_1, \ldots, \underline{X}_{p-1}]$, $\underline{\beta}' = [\beta_0, \beta_1, \ldots, \beta_{p-1}]$ and $\underline{\epsilon} \sim N(\underline{0}, \sigma_0^2 I_n)$, here I_n denotes the identity matrix of order $n \times n$. The model (1) having $p-1$ independent variables is considered as the true model. Any reduced model whose 'X matrix' has r columns is obtained by retaining any $r-1$ of the $p-1$ independent variables $X_1, \ldots, X_{p-1}$, where $2 \leq r \leq p-1$. For each r, $2 \leq r \leq p-1$, there are $k_r = \binom{p-1}{r-1}$ such models. These k_r reduced models of 'size' r are indexed arbitrarily with the indexing variable ℓ going from 1 to k_r. We will refer to a typical model as Model $M_{r\ell}$. If the i-th data is deleted, then the reduced Model $M_{r\ell}$ is denoted by $M_{r\ell(i)}$. A reduced model of size r can be written as

$$E(\underline{Y}) = X_{r\ell}\underline{\beta}_{r\ell}, \quad \ell = 1, 2, \ldots, k_r. \tag{18}$$

The reduced model for deleted i-th data is

$$E(\underline{Y}_{(i)}) = X_{r\ell(i)}\underline{\beta}_{r\ell(i)}, \quad \ell = 1, 2, \ldots, k_r. \tag{19}$$

It should be pointed out that all expectations and probabilities are calculated under the model (1).

Usually, we use the residual sum of squares to measure goodness of the fitted model for a random sample. Hence, the expected residual sum of squares is naturally considered as the measurement for the goodness of fit. Large values of this expectation are not

desirable. But, the estimate of the expectation is heavily influenced by the influential data. It is important to detect them, and consider them seriously. It should be first noted that our comparisons of models are made under the true model assumptions.

For any r, $2 \leq r \leq p-1$, the residual sum of squares $S_{r\ell}$ and $S_{r\ell(i)}$ for the reduced models $M_{r\ell}$ and $M_{r\ell(i)}$, $1 \leq \ell \leq k_r$, $i = 1, 2, \ldots, n$, are respectively as follows:

$$SS_{r\ell} = \underline{Y}' Q_{r\ell} \underline{Y}, \text{ and } S_{r\ell(i)} = \underline{Y}'_{(i)} Q_{r\ell(i)} \underline{Y}_{(i)} \tag{20}$$

where

$$Q_{r\ell} = [I_n - X_{r\ell}(X'_{r\ell} X_{r\ell})^{-1} X'_{r\ell}],$$

and

$$Q_{r\ell(i)} = [I_{n-1} - X_{r\ell(i)}(X'_{r\ell(i)} X_{r\ell(i)})^{-1} X'_{r\ell(i)}].$$

Also

$$\frac{SS_{r\ell}}{\sigma_0^2} \sim \chi^2\{n-r,\ \lambda_{r\ell}\},$$

and

$$\frac{SS_{r\ell(i)}}{\sigma_0^2} \sim \chi^2\{n-r-1, \lambda_{r\ell(i)}\} \tag{21}$$

where

$$\lambda_{r\ell} = (X\underline{\beta})' Q_{r\ell} (X\underline{\beta})/2\sigma_0^2,$$

and

$$\lambda_{r\ell(i)} = (X\underline{\beta})' Q_{r\ell(i)} (X\underline{\beta})/2\sigma_0^2.$$

We note that $Q_{r\ell}$ and $Q_{r\ell(i)}$ are idempotent and symmetric; thus it is positive semi-definite. Hence $\lambda_{r\ell}$ and $\lambda_{r\ell(i)}$ are nonnegative, but not zero, in general.

We have

$$E[SS_{r\ell}] = (n-r)\sigma_0^2 + 2\sigma_0^2 \lambda_{r\ell},$$

and

$$E[SS_{r\ell(i)}] = (n-r-1)\sigma_0^2 + 2\sigma_0^2 \lambda_{r\ell(i)}. \tag{22}$$

Since σ_0^2 is fixed, it is clear from (22) that $\lambda_{r\ell}$ and $\lambda_{r\ell(i)}$, for all i, should not be large for $M_{r\ell}$ as a good model.

Gupta and Huang (1988) have proposed some selection procedures for selecting good models based on $\lambda_{r\ell}$'s. Now, we are interested in studying the i-th data is deleted, how large influence for $\lambda_{r\ell(i)}$!

We have the unbiased estimate of $\lambda_{r\ell}$, σ_0^2, $\sigma_{0(i)}^2$ and $\lambda_{r\ell(i)}$ as follows:

$$\hat{\sigma}_0^2 = \frac{SS_{p1}}{n-p} = \frac{SSE}{n-p},\ \hat{\sigma}_{0(i)} = \frac{SS_{p1(i)}}{n-p-1} = \frac{SSE_{(i)}}{n-p-1}$$
$$\hat{\lambda}_{r\ell} = \frac{n-p}{2}\frac{SS_{r\ell}}{SS_{p1}} - \frac{n-r}{2} \tag{23}$$

and

$$\hat{\lambda}_{r\ell(i)} = \frac{n-p-1}{2} \cdot \frac{SS_{r\ell(i)}}{SS_{p1(i)}} - \frac{n-r-1}{2}.$$

Since

$$(X\underline{\beta} - X_{r\ell(i)}\hat{\underline{\beta}}_{r\ell(i)})'Q_{r\ell(i)}(X\underline{\beta} - X_{r\ell(i)}\hat{\underline{\beta}}_{r\ell(i)}) = (X\underline{\beta})'Q_{r\ell(i)}(X\underline{\beta}) = 2\sigma_0^2\lambda_{r\ell(i)}.$$

Hence, $\lambda_{r\ell(i)}$ also measures the influence of the i-th data on fitted values. We define the measurement of the influence for the i-th data as follows:

$$D_{r\ell(i)} = \lambda_{r\ell(i)} = \frac{(X\underline{\beta})'Q_{r\ell(i)}(X\underline{\beta})}{2\sigma_0^2} \tag{24}$$

We estimate $D_{r\ell(i)}$ in (24) as a statistic to measure the influence as follows:

$$\hat{D}_{r\ell(i)} = \hat{\lambda}_{r\ell(i)} = \frac{n-p-1}{2}\frac{SS_{r\ell(i)}}{SS_{p1(i)}} - \frac{n-r-1}{2}. \tag{25}$$

We can compute a constant d to satisfy the following equation:

$$\inf_{\lambda_{r\ell(i)} \geq \Delta} P\{\hat{D}_{r\ell(i)} \geq d\} = 1 - \alpha. \tag{26}$$

where $\Delta > 0$ and α are given.

We have in (25),

$$P\{\hat{D}_{r\ell(i)} \geq d | \lambda_{r\ell(i)} = \Delta\} = 1 - \alpha. \tag{27}$$

Since

$$V_{r\ell(i)} = \frac{[SS_{r\ell(i)} - SS_{p1(i)}]/(p-r)}{SS_{p1(i)}/(n-p-1)},$$

follows the noncentral F denoted as $F'(p-r,\ n-p-1;\ \lambda_{r\ell(i)})$ (cf. Graybill (1976)). Then, we have

$$[(d+\frac{n-r-1}{2})\frac{2}{n-p-1}-1]\frac{n-p-1}{p-r}=F'(p-r,\ n-p-1;\ \Delta) \tag{28}$$

From (28), we have

$$d=\frac{(n-p-1)}{2}\left\{\frac{(p-r)}{(n-p-1)}F'(p-r,\ n-p-1;\ \Delta)+1\right\}-\frac{n-r-1}{2}. \tag{29}$$

Patnaik (1949) provided an approximation to the noncentral F distribution (cf. Guenther (1979)) by the relation

$$\begin{aligned} F'(p-r,n-p-1;\Delta) &\approx \{[(p-r)+2\Delta]/p-r\}\,F(p^*,\ n-p-1), \\ \text{where } p^* &= [(p-r)+2\Delta]^2/[(p-r)+4\Delta]. \end{aligned} \tag{30}$$

Hence the constant d can be computed as follows:

$$d\approx\frac{(n-p-1)}{2}\left\{\frac{(p-r)+2\Delta}{n-p-1}F(p^*,\ n-p-1)+1\right\}-\frac{n-r-1}{2}. \tag{31}$$

We summarize the results as follows:

If the i-th data is deleted, and

$$\hat{D}_{r\ell(i)}\geq d, \tag{32}$$

then there exists an influential data in the reduced model $M_{r\ell}$.

A reduced model $M_{r\ell}$ is called an inferior model, if there is some i-th data which satisfies the condition (32), where i-th data is not an influential data in model (1). A method to select important independent regression variables is given in Gupta and Huang (1988).

To summarize, the selection processes are as follows:

Using Gupta and Huang (1988) procedure, we select some desirable reduced models denoted by T at stage 1. If any model in T is an inferior model, then we reject it. The set of the remaining models is denoted by T'. At stage 2, from the set T', we select the reduced model associated with the smallest total error using the statistic $\hat{\Gamma}_{r\ell}$, where $\hat{\Gamma}_{r\ell}=2\cdot\frac{n-p-2}{n-p}[2\hat{\lambda}_{r\ell}+(p-r)]-(2p-3r)$.

We shall take an example for the selection of influential data in Daniel and Wood (1980, p 234). The data were obtained in a laboratory study of the distillation properties of various crude oils with respect to their yield of gasoline. The four independent variables measured were:

X_1: crude oil gravity, $^\circ API$,

X_2: crude oil vapor pressure, psi,

X_3: crude oil ASTM 10% point, $^\circ F$,

X_4: gasoline ASTM end point, $^\circ F$,

Y: gasoline yield, as percentage of crude.

We fit the full model for the data as follows:

$$Y = -6.952 + 0.229X_1 + 0.553X_2 - 0.149X_3 + 0.155X_4.$$

$$R^2 = 0.96, \text{ Root MSE } = 2.231, \text{ C.V. } = \frac{\text{Root MSE}}{\overline{Y}} \times 100\% = 11.34\%.$$

where $\overline{Y}$ is the sample mean of Y_i's. We consider the reduced model as in Daniel and Wood (1980, p. 247):

$$Y = 70.84 - 0.212X_3 + 0.159(X_4 - 332) \tag{33}$$

$$R^2 = 0.95, \text{ Root MSE } = 2.426, C.V. = 12.338\%.$$

In the reduced model (33), there is no any influential data.

We have computed some values in TABLE III and TABLE IV to show some idea for the various statistics in the previous discussion.

We state the notation in the following table as follows:

Residual $= Y_i - \hat{Y}_i$,

RSTUDENT $= \frac{e_{(i)}\sqrt{1-h_{ii}}}{s_{(i)}}$

HAT DIAG $H = h_{ii}$,

$\hat{D}_{(i)} =$ (RSTUDENT)$^2 \times$ (HAT DIAG H)$/p$,

$C_i = \frac{1}{p} h_{ii} F(1,\ 32 - 5 - 1;\ 0.95)$,

$IF_{(i)} = \hat{D}_{(i)}/C_i$,

where $F(1,\ 26;\ 0.95) = (2.056)^2, n = 32$ and $p = 5$.

For the reduced model M_{31} in (33), and from (25), we have

$$\hat{D}_{31(29)} = \frac{32-5-1}{2} \times \frac{150.4016}{111.657} - \frac{32-3-1}{2}$$
$$= 3.51,$$

and from (31),

$$d = \frac{32-5-1}{2}\left\{\frac{(5-3)+2\Delta}{32-5-1}F(p^*, 32-5-1)+1\right\} - \frac{32-3-1}{2}$$

let $\Delta = 1.3$, $p^* = [(5-3)+2\Delta]^2/[(5-3)+4\Delta] = 2.94$, and let $\alpha = 0.05$,

$$F(2.94,\ 26;\ 0.95) \approx 3.00$$

we have $d \approx 5.9$. Thus, $\hat{D}_{31(29)} < d$. We have checked that $\hat{D}_{31(i)} < d$ for all $i = 1, \ldots, 32$. Hence, the reduced model (33) is reasonable to accept as a good model (not an inferior model). Note that the value of Δ can be chosen as in Gupta and Huang (1988). The value of Δ is the amount of bias for a reduced model in (26).

TABLE III

Obs.	Residual	RSTUDENT	HAT DIAG H
1	−1.8281	−0.8754	0.1340
2	−3.5300	−1.6607	0.0361
3	1.4166	0.6806	0.1494
4	−1.2104	−0.5915	0.1814
5	1.2999	0.6801	0.2829
6	2.8220	1.4356	0.1956
7	−0.6043	−0.2829	0.1170
8	−0.5769	−0.3001	0.2851
9	0.6229	0.3067	0.2014
10	−3.5804	−1.7239	0.0730
11	−0.4090	−0.1844	0.0495
12	0.3604	0.1680	0.1114
13	2.8245	1.4179	0.1755
14	0.2360	0.1117	0.1390
15	−0.3763	−0.1734	0.0901
16	−1.3024	−0.6501	0.2134
17	1.0405	0.5009	0.1598
18	−2.8326	−1.3429	0.0823
19	3.3608	1.6105	0.0763
20	2.9376	1.3887	0.0729
21	−1.7290	−0.8222	0.1248
22	−1.4534	−0.7120	0.1808
23	−2.1059	−1.0936	0.2519
24	−2.8999	−1.3971	0.1067
25	1.9063	0.9332	0.1682
26	0.2721	0.1356	0.2232
27	1.4451	0.7112	0.1881
28	−2.1919	−1.0538	0.1299
29	4.6214	2.2937	0.0586
30	−0.0500	−0.0261	0.2897
31	−0.1696	−0.0809	0.1521
32	1.6838	0.8975	0.3000

TABLE IV

Obs.	$\hat{D}_{(i)}$	C_i	$SSE_{(i)}$	$IF_{(i)}$
1	0.021	0.113	130.6	0.181
2	0.020	0.031	121.5	0.652
3	0.014	0.126	132.0	0.109
4	0.013	0.153	132.6	0.083
5	0.026	0.239	132.0	0.109
6	0.081	0.165	124.5	0.488
7	0.002	0.099	134.0	0.019
8	0.005	0.241	133.9	0.021
9	0.004	0.170	133.9	0.022
10	0.04	0.062	120.7	0.703
11	0.000	0.042	134.2	0.008
12	0.001	0.094	134.3	0.007
13	0.071	0.148	124.7	0.476
14	0.000	0.118	124.7	0.003
15	0.001	0.076	134.2	0.007
16	0.018	0.180	132.2	0.100
17	0.008	0.135	133.1	0.059
18	0.030	0.070	125.7	0.427
19	0.040	0.065	122.2	0.614
20	0.028	0.062	125.5	0.456
21	0.017	0.106	131.0	0.160
22	0.018	0.153	131.8	0.120
23	0.060	0.213	128.5	0.283
24	0.042	0.090	124.9	0.462
25	0.029	0.142	130.1	0.206
26	0.001	0.189	134.3	0.004
27	0.019	0.159	131.8	0.120
28	0.029	0.110	128.9	0.263
*29	0.062	0.050	111.7	1.245
30	0.000	0.245	134.4	0.000
31	0.000	0.129	134.4	0.002
32	0.048	0.254	130.4	0.191

Remark: * denotes the influential data.

$$\underline{\hat{\beta}}_{(29)} = \begin{pmatrix} -8.05 \\ 0.25 \\ 0.52 \\ -0.15 \\ 0.15 \end{pmatrix}, \quad \underline{\hat{\beta}} = \begin{pmatrix} -6.95 \\ 0.23 \\ 0.55 \\ -0.15 \\ 0.15 \end{pmatrix}.$$

The effect on $\underline{\hat{\beta}}$ of deleting the observation number 29, is shown in the above two values of $\underline{\hat{\beta}}_{(29)}$ and $\underline{\hat{\beta}}$. The big change takes place in $\hat{\beta}_0$.

References

[1] Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics* **19**, 15–18.

[2] Cook, R. D. and Weisberg, S. (1980). Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics* **22**, 495–508.

[3] Gray, J. B. and Ling, R. F. (1984). K-Clustering as a detection tool for influential subsets in regression. *Technometrics* **26**, 305–318.

[4] Graybill, F. A. (1976). *Theory and Application of the Linear Model*. Duxbury Press, North Scituate, MA.

[5] Guenther, W. C. (1979). The use of noncentral F approximations for calculation of power and sample size. *Amer. Statist.* **33** (**4**) , 209–210.

[6] Gupta, S. S. and Huang, D. Y. (1988). Selecting Important Independent Variables in Linear Regression Models. *JSPI* **20**, 155–167.

[7] Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data*. 2nd edition. John Wiley & Sons, New York.

[8] Patnaik, P. B. (1949). The noncentral chi-squared and F distributions and their applications. *Biometrika* **36**, 202–232.

[9] Welsch, R. E. (1982). Influence functions and regression diagnostics. Modern Data Analysis (Launer, R. L. and Siegel, A. F.). New York: Academic Press.

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release, distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) Technical Report #89-28C	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Purdue University 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) Department of Statistics West Lafayette, IN 47907	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Office of Naval Research 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER NSF Grants DMS-8606964, DMS-8702620 N00014-88-K-0170
8c. ADDRESS (City, State, and ZIP Code) Arlington, VA 22217-5000	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
ON DETECTING INFLUENTIAL DATA AND SELECTING REGRESSION VARIABLES (Unclassified)

12. PERSONAL AUTHOR(S)
Shanti S. Gupta and Deng-Yuan Huang

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical	FROM _____ TO _____	October 1989	17

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Linear Model, Influential data, Selection criteria, Inferior model

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

We consider a linear regression model. We attempt to measure the influence on the residual for the full model and the reduced model. A criterion to select important independent regression variables is also derived using the influence diagnostics. This criterion turns out to be the same as the one proposed by Gupta and Huang (1988).

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS	Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Shanti S. Gupta	22b. TELEPHONE (Include Area Code) (317)494-6031 / 22c. OFFICE SYMBOL

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED